基于称名反应模型的 CD-CAT 选题方法比较

张杰 [1] 罗照盛 [**1] 喻晓锋 [**1] 秦春影 [2]

([1]江西师范大学心理学院，南昌，330022)
([2]南昌师范学院数学与信息科学学院，南昌，330032)

摘 要 当前大多数 CD-CAT 有关的研究都是基于 0-1 计分的数据资料展开的，而在实际的教育与心理测验中，还包含大量称名反应数据。本文基于称名反应认知诊断模型（NR-cRUM）开发了适用于称名反应数据的 CD-CAT（以下简称 NCD-CAT），并将 7 种 0-1 计分 CD-CAT 的选题方法引入 NCD-CAT 中。比较不同条件下，不同选题方法对被试判准率和测验效率的影响。结果表明 NR_PWCDI、NR_MPWKL 等 PWKL 系新方法和 NR_SHE/MI 方法能较好地适用于 NCD-CAT，且在大多数条件下优于基线方法 NR_PWKL。研究拓展了称名多级计分 CD-CAT 的选题方法。

关键词 CD-CAT；称名反应；NR-cRUM；选题方法；多项选择题

1. 引言

认知诊断计算机化自适应测验（CD-CAT; Cheng, 2009）结合了认知诊断评价（CDA; Leighton & Gierl, 2007; von Davier & Lee, 2019）和计算机化自适应测验（CAT）两者的优势（罗照盛 等, 2015; Yu et al., 2019）。根据计分方式的不同，CD-CAT 可分为 0-1 计分和多级计分的 CD-CAT。目前多数的 CD-CAT 研究都是基于 0-1 计分数据展开的。

然而在实际的教育和心理测验中，也存在很多的多级计分项目（刘拓 等, 2015; 王晓庆 等, 2016）。根据作答类别（response categories）之间有无顺序或等级，多级计分项目又可分为称名多级（nominal polytomous）和顺序多级（ordinal polytomous）计分。其中，称名多级计分数据常见于多项选择题（multiple choice items, MCI）、个性或态度量表中只反映不同倾向并无明确正确答案的题目中，是指多个作答类别之间相对独立、无顺序或等级之分的数据。称名多级计分数据可认为是最一般、测量级别最低的数据类型，基于顺序或等级的多级计分以及 0-1 计分数据均可看成是称名多级计分的特例（Mellenbergh, 1995）。

为了能分析并提取称名反应数据中的信息，研究者们开发了相应的称名反应类模型。在 IRT 的框架下，Bock（1972）开发了称名反应模型（nominal response model, NRM），并将该模型用于分析称名计分的多项选择题。结果表明，NRM 能够利用错误选项中的信息，其能力估计精度显著高于普通 0-1 计分 IRT 模型的估计精度；对于中低能力被试而言，NRM 模型的能力估计精度能够达到 2 倍测验长度的 0-1 计分 IRT 模型的估计精度。Wang 等（2017）将 NRM 用于构建可修改答案的 CAT，利用 NRM 将被试的第一次作答及后续的修改作答均纳入到临时的和最终的能力估计中，从而提供更多的选题信息和更准确的能力估计值。该可修改答案的 CAT 方案正是利用了称名反应模型中各作答类别是称名计分这一特点。进一步，Wang 等（2019）对基于 NRM 的可修改答案的 CAT 方案从理论上讨论了其可行性。

在认知诊断的框架下，Templin 等（2008）和 de la Torre（2009）开发了称名反应认知诊断模型。依据传统 0-1 计分的认知诊断模型，被试只能被归为掌握组和未掌握组两个类别。而在称名反应认知诊断模型中，被试可以被分成更多的类别，从而能够提高对被试的

*本研究得到江西省教育科学十四五规划 2021 课题（21YB027）的资助
**通讯作者：喻晓锋，E-mail：xyu6@jxnu.edu.cn；罗照盛：luozs@jxnu.edu.cn

分类精度。Templin 等（2008）将对数线性认知诊断模型（log-linear CDM, LCDM）进行称名多级化拓展，开发了 NR-DM（nominal response diagnostic model）及其缩减模型 NR-cRUM（nominal response compensatory reparameterized unified model，关于该模型的详细介绍见附录 A）。与 0-1 计分 cRUM 相比，使用称名多级计分的 NR-cRUM 对被试的判准率更高。de la Torre（2009）将 DINA 模型进行称名多级化拓展，拓展后的 MC-DINA 模型的判准率显著高于 DINA 模型的判准率，原因是 MC-DINA 模型利用了干扰项中的诊断信息。郭磊和周文杰（2021）提出一类非参数化的诊断干扰项中信息的方法。

在查阅了国内外相关文献之后发现，仅有 Yigit 等（2019）开发了基于称名反应模型的 CD-CAT。在该研究中，作者仅仅评估了一种选题方法的效率（通过与 0-1 计分的 CD-CAT 相比），使用的称名反应认知诊断模型是 MC-DINA 模型。MC-DINA 模型是一种非补偿模型，只适用于属性间的非补偿情形，且由于参数数量有限，模型不能直接解释每个属性与每个作答类别之间的关系，从而限制了模型参数解释的一般性（罗照盛 等, 2020）。由于全模型的 NR-DM 参数较多，估计这些参数需要很大的样本量，所以其缩减形式 NR-cRUM 实用性更大（李瑜, 2014; Templin et al., 2008），故本文采用 NR-cRUM 模型来开发多级计分 CD-CAT，将 7 种常见的选题方法拓展到适用于 NR-cRUM 的诊断测验中，并对它们在不同实验条件下的效率和精度进行综合比较。

2. 基于称名反应的 CD-CAT

2.1 初始题

在 CD-CAT 的初始题阶段，研究者提出随机选取的方法；或给被试随机指定一种属性掌握模式（KS），再利用选题方法选出相应的题（高椿雷 等, 2017; Yu et al., 2019）。Zheng 和 Chang（2016）提出的 PWCDI 和 PWACDI 选题法可以完全排除初始阶段 KS 估计不稳定的问题，天然地选出符合“T 阵法”（涂冬波 等, 2013）初始选题法要求的题目。为了各选题法之间比较的公平，本研究拟采用随机选取 1 题并在各选题法中都使用这一题作为初始题的方法。

2.2 适用于 NCD-CAT 的选题方法

在 0-1 计分的 CD-CAT 中，研究者们提出了多种选题方法（郭磊 等, 2016; 李佳 等, 2021; 罗照盛 等, 2015; Cheng, 2009; Guo & Zheng, 2019; Kaplan et al., 2015; Wang, 2013; Yu et al., 2019; Zheng & Chang, 2016）。现有的多级计分 CD-CAT 选题方法（Gao et al., 2020）主要从 0-1 计分的 CD-CAT 拓展而来，本研究沿用这一方法，将传统 CD-CAT 中效果较好的几种选题方法拓展至 NCD-CAT 中。

假设有 K 个属性将被试分为 $C = 2^K$ 个潜在类，第 j 题有 $b_j + 1$ 个选项（1 个正确选项，b_j 个干扰项）。则本研究所涉及的 NCD-CAT 选题方法介绍如下。

2.2.1 KL 信息矩阵及其变式

在介绍选题方法之前，先简要介绍 KL 信息矩阵。KL 信息矩阵又称 D 矩阵（Henson & Douglas, 2005），是一个 $2^K \times 2^K$ 的矩阵（K 为属性个数），它的每个元素是（给定作答反应条件下）两个 KS 之间的 KL 距离期望值。计算公式如下：

$$D_{juv} = E_{\alpha_u}\left[\log\left(\frac{P_{\alpha_u}(X_j)}{P_{\alpha_v}(X_j)}\right)\right] \quad (1)$$

其中，D_{juv} 是给定题目 j 的作答为 X_j 时，$\boldsymbol{\alpha_u}$ 和 $\boldsymbol{\alpha_v}$ 之间 KL 距离的期望值；当 X 是 0-1 计

分时，D_{juv} 计算公式如下：

$$D_{juv} = E_{\alpha_u}\left[\log\left(\frac{P_{\alpha_u}(X_j)}{P_{\alpha_v}(X_j)}\right)\right] = \sum_{x_j=0}^{1} P_{\alpha_u}(x_j)\log\left[\frac{P_{\alpha_u}(x_j)}{P_{\alpha_v}(x_j)}\right] \tag{2}$$

在 NCD-CAT 中，X 是称名计分的。相应地计算 D 矩阵（此时记为 NR_D）时，应该按照每个作答类别的概率求期望值，即：

$$NR_D_{juv} = \sum_{x_j=0}^{b_j} P_{\alpha_u}(x_j)\log\left[\frac{P_{\alpha_u}(x_j)}{P_{\alpha_v}(x_j)}\right] \tag{3}$$

NR_D_{juv} 矩阵包含了各个作答类别区分不同 KS 的能力信息，将会比传统 0-1 计分的 D 矩阵包含更多的信息。本研究中的 PWKL 系列选题法（NR_PWKL, NR_PWCDI, NR_PWACDI, NR_MPWKL）都是以 NR_D 矩阵为基础，再结合相应 0-1 计分 CD-CAT 选题法的思想拓展而来，下文相同情况不再赘述。

题目水平的 D 矩阵能够表示该题的信息量，有研究者对 D 矩阵进行了不同形式的加权求和，得到 CDI 和 ACDI 指标（Henson & Douglas, 2005; Henson et al., 2008）。其中 CDI 是 D 矩阵中的所有元素按两 KS 之间的海明距离（hamming distance）进行加权求平均的结果，而 ACDI 是将 D 矩阵中海明距离为 1 的元素相加后求平均值，计算公式如下：

$$CDI_j = \frac{\sum_{u\neq v} h(\boldsymbol{\alpha}_u,\boldsymbol{\alpha}_v)^{-1}\cdot D_{juv}}{\sum_{u\neq v} h(\boldsymbol{\alpha}_u,\boldsymbol{\alpha}_v)^{-1}} \tag{4}$$

$$ACDI_j = \sum_{k=1}^{K} ACDI_{jk} = \sum_{k=1}^{K}\frac{1}{2^K}\sum_{all\ relevant\ cells} D_{juv} \tag{5}$$

其中，all relevant cells 指 D 矩阵中两 KS 的海明距离为 1 的单元格。

2.2.2 NR_PWKL

NCD-CAT 的 PWKL 指标（以下称 NR_PWKL）是 PWKL（posterior-weighted KL）选题法（Cheng, 2009）的称名多级化拓展。PWKL 选题法对 D 矩阵的每个元素按每种 KS 的后验概率加权求和，得到 PWKL 指标。而 NR_PWKL 选题法对 NR_D 矩阵中每个元素进行相同加权求和，计算公式如下：

$$NR_PWKL_j(\hat{\boldsymbol{\alpha}}_i) = \sum_{c=1}^{2^K}\left\{\left[\sum_{y=0}^{b_j}\log\left(\frac{P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_c)}\right)P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)\right]\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)\right\} \tag{6}$$

其中，$\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)$ 是观察到作答向量为 $\boldsymbol{y}_t$时，被试的属性掌握模式是 $\boldsymbol{\alpha}_c$ 的后验概率。

2.2.3 NR_PWCDI 和 NR_PWACDI

Zheng 和 Chang（2016）遵循 PWKL 的思想，在 D 矩阵的行和列同时加入 KS 的后验概率，构造出 PWD 矩阵（posterior-weighted D matrix），再按照 CDI 和 ACDI 指标加权的思想对 PWD 矩阵进行不同加权处理，得到 PWCDI 和 PWACDI 指标。类似地，本文在 NR_D 矩阵的行和列同时加入后验概率得到 NR_PWD 矩阵，即公式（7）。再根据两种不同的加权思想将 NR_PWD 矩阵和 NR_D 矩阵加权求均值后得到 NR_PWCDI 和 NR_PWACDI 指标，即公式（8），（9）。

$$NR_PWD_{jic} = \pi(\boldsymbol{\alpha}_i)\times\pi(\boldsymbol{\alpha}_c)\times\left[\sum_{y=0}^{b_j}\log\left(\frac{P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_c)}\right)P(Y_{ij}=y|\hat{\boldsymbol{\alpha}}_i)\right] \tag{7}$$

$$NR_PWCDI_j = \frac{\sum_{i\neq c} h(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_c)^{-1}\cdot NR_PWD_{jic}}{\sum_{i\neq c} h(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_c)^{-1}} \tag{8}$$

$$NR_PWACDI_j = \frac{1}{2^K}\sum_{all\ relevant\ cells} NR_D_{juv} \tag{9}$$

同理，all relevant cells 指 NR_D 矩阵中两 KS 的海明距离为 1 的单元格。

2.2.4 NR_MPWKL

Kaplan 等（2015）、Zheng 和 Chang（2016）的研究都指出，在 KL 系列方法中，若仅使用当前 KS 估计值的后验概率加权，不利于选出最合适的题。因为当测验较短时，当前 KS 的估计值通常都不太准确。除上述 PWCDI 类方法外，Kaplan 等（2015）的 MPWKL 方法也能很好地解决这个问题。MPWKL 方法对 PWKL 指标按照各 KS 的后验概率再进行加权求和。本研究对 NR_PWKL 指标按照 KS 的后验概率再次加权求和，得到 NR_MPWKL 指标，计算公式如下：

$$NR_MPWKL_j(\widehat{\boldsymbol{\alpha}}_i) = \sum_{d=1}^{2^K}\left\{\sum_{c=1}^{2^K}\left[\sum_{y=0}^{b_j}\log\left(\frac{P(Y_{ij}=y|\widehat{\boldsymbol{\alpha}}_i)}{P(Y_{ij}=y|\boldsymbol{\alpha}_c)}\right)P(Y_{ij}=y|\widehat{\boldsymbol{\alpha}}_i)\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)\right]\pi(\boldsymbol{\alpha}_d|\boldsymbol{y}_t)\right\} \tag{10}$$

与 0-1 计分一样，上述 4 种 PWKL 系新方法也是选择相应指标最大的题目。

2.2.5 NR_SHE

将香农熵（Shannon Entropy, SHE）应用于选题方法中，通过选择题库中期望香农熵最小的题目，使估计的 KS 后验概率分布的不确定性最小。与 SHE 不同的是，计算期望香农熵 NR_SHE 时，应该按照每个作答类别（ b_j+1个）的概率求香农熵的期望值，而不是仅仅考虑 0 和 1 两个作答类别。计算公式如下：

$$NR_SHE = \sum_{y=0}^{b_j}\left[\sum_{c=1}^{2^K}\left(\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t,Y_{t+1}=y)\log\frac{1}{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t,Y_{t+1}=y)}\right)P_r(Y_{t+1}=y|\boldsymbol{y}_t)\right] \tag{11}$$

其中 $Pr(Y_{t+1}=y|\boldsymbol{y}_t)$ 是观察到前 t 次作答的向量为 $\boldsymbol{y}_t$，第 $t+1$ 题的得分为 y 的条件概率，其计算公式如下：

$$\sum_{c=1}^{2^K} P(Y_{t+1}=y|\boldsymbol{\alpha}_c)\,\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t) \tag{12}$$

其中 $\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)$ 是被试完成 t 道题目后，被试的属性掌握模式是 $\boldsymbol{\alpha}_c$ 的后验概率。

2.2.6 NR_MI

互信息（Mutual Information, MI）选题法是 Wang（2013）提出的一种更适用于短测验的选题方法。类似地，求 NR_MI 时，也应该分别按照 b_j+1个作答类别的概率求 MI 的期望，计算公式如下：

$$NR_MI = \sum_{y=0}^{b_j}\left[\sum_{c=1}^{2^K}\left(\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t,Y_{t+1}=y)\log\frac{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t,Y_{t+1}=y)}{\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)}\right)\sum_{c=1}^{2^K}P(Y_{t+1}=y|\boldsymbol{\alpha}_c)\,\pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)\right] \tag{13}$$

该方法选择题库中 NR_MI 指标最大的题目。

2.2.7 NR_GDI

Kaplan 等（2015）将 GDI（G-DINA discrimination index, GDI）指标用于 CD-CAT 的选题中，其中 $\bar{P}_j$ 是除第一个选项外，其他选项（ b_j 个）的平均得分。

$$NR_GDI = \sum_{c=1}^{2^K} \pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)\left[\sum_{y=1}^{b_j} P(Y_{ij} = y|\boldsymbol{\alpha}_c) - \bar{P}_j\right] \tag{14}$$

$$\bar{P}_j = \sum_{c=1}^{2^K} \pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t) \sum_{y=1}^{b_j} P(Y_{ij} = y|\boldsymbol{\alpha}_c) \tag{15}$$

与 PWKL 系列指标一样，GDI 指标越大，表示区分不同 KS 的能力越强。所以，NR_GDI 方法选择题库中 NR_GDI 指标最大的题目。

2.3 曝光控制

Zheng 和 Wang（2017）基于计算机领域的二分搜索算法，开发了 SDBS 与 DBS 方法。通过与已有的控制题目曝光的方法（RP, RT, SHTVOR 等）相比，新方法能够更高效地处理好测验准确率与曝光控制之间的权衡问题。本研究聚焦于各选题方法在被试分类精度和测验效率上的表现，故没有考虑曝光控制的问题。

2.4 参数估计方法

在 CD-CAT 中有三种参数估计方法，分别是 MLE, MAP, EAP（Huebner & Wang, 2011）。其中 EAP 计算的是被试属性掌握模式的期望后验概率，是在给定作答为 $\boldsymbol{y}_t$ 的条件下，将所有可能的 KS 与其对应的后验概率相乘再求期望值（公式 16），最后再进行二分取值转换。（涂冬波等, 2017）。本研究采用的是 EAP 方法。

$$\hat{p}_{ik} = \sum_{c=1}^{2^K} \pi(\boldsymbol{\alpha}_c|\boldsymbol{y}_t)\alpha_{ck} \tag{16}$$

2.5 终止策略

目前 CD-CAT 的终止策略主要有定长和变长两类。Guo 和 Zheng（2019）指出：变长终止策略中的 Tatsuoka 规则和双标准规则在属性个数不同时，存在不稳定问题；并从信息论的视角提出了变长终止策略的新方法。本研究的重点是选题方法的比较，因此终止策略仅考虑定长和最大后验概率达到某一固定精度的变长策略（Yu et al., 2019）。

3. 模拟研究

为考察和比较 NCD-CAT 不同选题方法的性能，本研究开展了两项实验。

3.1 实验一：定长 NCD-CAT

实验一是一个 $2 \times 4 \times 7$ 的三因素完全随机实验设计。自变量分别是题目质量、题长、选题方法。其中题目质量有高、低两个水平，题长有 5, 10, 15, 20 四个水平，选题方法有 NR_PWKL, NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, NR_MI, NR_GDI 7 种方法。具体实验过程描述如下：

3.1.1 数据模拟

被试方面，本实验假设 5 个独立属性共 32 种属性掌握模式的被试在人群中均匀分布，模拟生成 3200 名被试。

题库方面，本实验采用 Ma 和 de la Torre（2016）类似的模拟方法生成高、低质量的题库（各 600 题），Q 矩阵采用 de la Torre（2009）的 Q 矩阵（见附录 B）。对于高（或低）质量的题目，掌握该题目所考察的属性的被试选择正确答案的概率为 0.8（或 0.6）；被试选择错误答案的概率按均匀分布模拟。

模拟作答方面，先在（0, 1）中生成一个均匀分布的随机数，然后比较这个随机数落在 A, B, C, D 哪一个累积作答概率区间内，就选择这个选项作为答案。(例如，当某被试选择各选项的概率依次是 0.1, 0.3, 0.5, 0.1，则累积作答概率分布为 0.1, 0.4, 0.9, 1，如果随机数为 0.63，则模拟该被试选择第三个选项)。

3.1.2 评价指标

定长 NCD-CAT 的评价指标包括模式判准率（pattern match ratio, PMR）、χ^2 和测验重叠率（test overlap rate, TOR）。各评价指标的详细说明见附录 C。

3.1.3 实验一结果

实验一结果在表 1，附录 D：表 D-1，图 D-1，图 D-2 中。

表 1 七种选题方法的模式判准率、与 NR_PWKL 比较的差值

题长	方法 (NR-)	高质量		低质量	
		PMR	**Difference**	**PMR**	**Difference**
5	PWKL	0.624		0.335	
	PWCDI	0.711	0.087	0.367	0.032
	PWACDI	0.714	0.090	**0.373**	0.038
	MPWKL	0.704	0.080	0.357	0.022
	SHE	**0.736**	**0.112**	0.352	0.017
	MI	**0.736**	**0.112**	0.352	0.017
	GDI	0.672	0.048	0.322	-0.013
10	PWKL	0.917		0.621	
	PWCDI	0.948	0.031	0.643	0.022
	PWACDI	0.948	0.031	0.633	0.012
	MPWKL	0.948	0.031	0.641	0.020
	SHE	**0.951**	0.034	**0.646**	0.025
	MI	**0.951**	0.034	**0.646**	0.025
	GDI	0.948	0.031	0.607	-0.014
15	PWKL	0.988		0.789	
	PWCDI	0.993	0.005	0.798	0.009
	PWACDI	0.993	0.005	**0.808**	0.019
	MPWKL	**0.994**	0.006	0.803	0.014
	SHE	0.993	0.005	**0.808**	0.019
	MI	0.993	0.005	**0.808**	0.019
	GDI	0.993	0.005	0.772	-0.017
20	PWKL	0.999		0.888	
	PWCDI	0.999	0	0.892	0.004
	PWACDI	0.999	0	0.897	0.009
	MPWKL	**1**	0.001	0.895	0.007
	SHE	0.999	0	**0.902**	0.014
	MI	0.999	0	**0.902**	0.014
	GDI	0.998	-0.001	0.874	-0.014

整体趋势（表 1）：

（1）随着题目质量由低变高，各选题方法的 PMR 都明显提高，提升效果在短测验（5 题、10 题）中尤为明显。例如当题长为 5 题时，各选题法的 PMR 几乎提升了一倍。(2) 随着题长的增加，各选题法的 PMR 都有提高，提高的程度因题长和题目质量而异。具体而言，当题长由 5 题增加到 10 题时，不管题目质量高或低，它们的 PMR 都有超过 20%的提升；当题长由 10 题增加到 15 或 20 题时，低质量题目 NCD-CAT 的 PMR 仍有接近或超过 10%的提升。相较而言，提高题目质量比增加题长更能提高 NCD-CAT 的判准率。

各选题方法与基线方法 NR_PWKL 的比较（表 1）：

（1）NR_PWKL 系新方法（NR_PWCDI, NR_PWACDI, NR_MPWKL）和 NR_SHE/MI 方法的 PMR 在所有条件下都高于或等于 NR_PWKL 法，尤其是在题长为 5 和高质量题目条件下；（2）随着题长的增加，其余方法相较于 NR_PWKL 法的 PMR 优势不断减小；这一趋势与 Zheng 和 Chang（2016）研究中 PWCDI, PWACDI 和 MPWKL 方法的变化一致；

NR_PWCDI 和 NR_PWACDI 与 NR_MPWKL 和 NR_SHE/MI 方法的比较（表 1）：

（1）NR_PWCDI 和 NR_PWACDI 方法与 NR_MPWKL 方法表现非常接近，与 Zheng 和 Chang（2016）的结果一致，因为这类方法都是基于 PWKL 的改进；（2）香农熵类方法（SHE/MI）的 PMR 高于或等于 NR_PWKL 系方法；Wang（2013）指出，MI 选题法是一种在短测验中表现较好的方法。

NR_GDI 方法在低质量题目条件下是 7 种方法中表现最差的，而在高质量题目条件下略微好于基线方法 NR_PWKL。

测验重叠率（TOR）和卡方值（χ^2）（见附录表 D-1）：

（1）NR_GDI 的 TOR 和卡方值最大，题库安全性较差；NR_MPWKL 次之；（2）除 NR_GDI 方法外，随着题目质量的提升，同等题长条件下其它方法的 TOR 均下降；随着题长的增加，同等题目质量条件下其它方法的 TOR 均上升；上述两点与 Gao 等（2020）的研究结果一致；（3）NR_PWCDI 和 NR_PWACDI 方法的 TOR 和卡方值大于 NR_PWKL 法，但小于 NR_MPWKL 法；（4）相较而言，NR_SHE/MI 方法的 TOR 和卡方值在不同条件下变化较平稳。

不同类型 KS 被试的判准率（见附录图 D-1，图 D-2）：

实验一还进一步分析了各选题方法对不同类型 KS 被试的判准率。由于所有题长条件下其变化趋势一致，故仅列出题长为 10 这一种情况。结果表明：

（1）题目质量对判准率的影响较大。题目质量由高到低，各方法对不同 KS 的 PMR 整体下降了 30%左右；（2）当题目质量高时，NR_PWKL 对不同 KS 的 PMR 较低；而当题目质量低时，NR_GDI 取代 NR_PWKL 成为 PMR 最低的选题方法；上述两点在表 1 中也得到印证。

3.2 实验二：变长 NCD-CAT

实验二是 2 × 3 × 7 的三因素完全随机实验。实验二将实验一的题长因素替换成最大后验概率的因素，其余自变量、数据模拟和参数估计方法均与实验一相同。实验二中最大后验概率分别是 0.8, 0.85, 0.9；终止条件除了最大后验概率，还增设一个最大题长为 20 题的条件。

实验二的评价指标是 PMR 和测验效率，后者主要体现在最大、最小、平均题长及题长的标准差指标上。

3.2.1 实验二结果

实验二结果在表 2，附录 D：表 D-2，图 D-3，图 D-4 中。

各选题方法的平均题长以及与 NR_PWKL 的比较（表 2）：

（1）NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, NR_MI 选题法在所有实验条件下，都比 NR_PWKL 的平均题长更短；尤其题目质量高时优势更明显，差值大于 0.738 题；（2）NR_GDI 方法则受题目质量影响，在高（或低）质量条件下题长小（或大）于 NR_PWKL；（3）NR_PWCDI, NR_PWACDI 和 NR_MPWKL 三种方法在所有变长条件下表现接近；与 NR_SHE/MI 法比较方面，当终止规则较宽松或题目质量较好时，NR_SHE/MI 的题长更短；反之前面三种方法的题长更短。

题长的其他描述统计量汇总见附录表 D-2。

表 2 各选题方法的平均题长以及与 NR_PWKL 比较的差值

终止规则	方法 (NR-)	高质量		低质量	
		Mean	Difference	Mean	Difference
	PWKL	7.108		13.6	
	PWCDI	6.249	0.859	**13.258**	0.314
	PWACDI	6.251	0.857	13.301	0.271
0.8	MPWKL	6.262	0.846	13.268	0.299
	SHE	**6.119**	0.989	13.297	0.252
	MI	**6.119**	0.989	13.297	0.252
	GDI	6.492	0.616	14.227	-0.529
	PWKL	7.72		15.009	
	PWCDI	6.822	0.897	14.702	0.284
	PWACDI	6.823	0.896	14.694	0.285
0.85	MPWKL	6.83	0.889	**14.648**	0.304
	SHE	**6.766**	0.953	14.781	0.194
	MI	**6.766**	0.953	14.781	0.194
	GDI	7.056	0.663	15.737	-0.558
	PWKL	8.394		16.884	
	PWCDI	**7.549**	0.844	16.588	0.24
	PWACDI	7.585	0.808	16.535	0.252
0.9	MPWKL	7.563	0.83	**16.52**	0.274
	SHE	7.655	0.738	16.693	0.148
	MI	7.655	0.738	16.693	0.148
	GDI	7.884	0.509	17.653	-0.505

不同类型 KS 被试的平均题长（见附录图 D-3，图 D-4）：

实验二也进一步分析了不同 KS 类型的被试在各选题方法下的测验效率。由于所有条件下其变化趋势一致，故仅列出最大后验概率为 0.85 这一种情况。结果表明：

（1）NR_PWKL 对几乎所有 KS 被试的题长最长，第二长的方法是 NR_GDI；（2）NR_SHE/MI 方法对所有 KS 被试的题长最短；这两点发现与表 2 的结果是一致的。（3）随着被试掌握的属性个数增加，所有选题方法的平均题长呈现阶跃式下降，下降的拐点都出现在属性个数增加的地方；即在低质量题目条件下，能够通过多选择一些考察多个属性的题目给被试，以使得测验效率有较大提升；（4）在低质量题目条件下对于所有 KS 的被试，NR_GDI 方法的平均题长最长，测验效率最低；NR_PWKL 次之，其余方法差异不明显；这与表 2 的发现也是一致的。

4. 总结与展望

认知诊断评价以诊断被试的认知优势和劣势见长，而 CAT 的优点是高效、精准地测量被试的能力。结合了两者优势的 CD-CAT 将会给教育工作者带来极大的便利。然而目前 CD-CAT 应用于实践还不多，主要是还有不少问题亟待解决。本研究从 CD-CAT 无法充分利用称名反应数据中的信息入手，开发 NCD-CAT 选题方法。研究一比较了各选题方法在不同题长和不同题目质量条件下的 NCD-CAT 表现，研究二对变长条件下各选题方法的表现进行了探究。结果表明：（1）NR_PWCDI, NR_PWACDI 和 NR_MPWKL 方法在各实验条件下表现近似，且一致优于 NR_PWKL 方法；NR_SHE/MI 方法与上述 3 种 NR_PWKL 系新方法相比，在短测验时优于它们，但优势不大；（2）题目质量对测验判准率和测验效率的影响较大，在实际应用时应该挑选高质量的题目进入题库。研究拓展了称名多级计分 CD-CAT 的选题方法。

本文提出的 NCD-CAT 适用于称名反应数据，虽然可用于分析多选题各选项中的信息，

但其应用范围仍然受到限制。后续研究有以下几个方面值得进一步探讨。(1)本文的 NCD-CAT 是以一个特殊的认知诊断模型为基础开发的多级计分 CD-CAT，未来可以考虑使用一般的多级计分认知诊断模型（Gao et al., 2021)。(2)本文 NCD-CAT 的选题方法都是由 0-1 计分的 CD-CAT 拓展而来，未来可考虑开发针对多级计分 CD-CAT 特点的选题方法。例如 Yigit 等（2019）使用 JSD（Jensen Shannon divergence）指标作为基于 MC-DINA 模型的 CD-CAT 选题方法，未来可将 JSD 法应用于 NCD-CAT 中，以考察该选题方法的效果。(3)本研究的实验条件较为理想，还有很多实际的问题没有考虑进来，例如曝光控制、初始阶段的选题法、其他变长终止规则等，相关议题的最新研究成果值得借鉴。例如，未来可考虑 Zheng 和 Wang（2017）开发的 SDBS 与 DBS 方法，以处理好测验准确率与曝光控制之间的权衡问题；可利用 Zheng 和 Chang（2016）提出的 PWCDI 和 PWACDI 选题法选出更合适的初始题；可结合 Guo 和 Zheng（2019）提出的变长 CD-CAT 终止策略的新方法，检验新方法在不同 CDM 中的稳定性。(4)基于称名反应数据的 CDM 最自然的应用是用于提取多选题干扰项中的诊断信息，未来可考虑比较不同称名反应认知诊断模型挖掘干扰项信息的效果。(5)本文研究的测验数据是基于 0-1 属性、多级计分，该类型数据能够提供更丰富的诊断信息。然而在 0-1 计分的框架下，也可通过属性多级化的方式来丰富测验数据中的信息，尤其是当考虑多级属性之间的顺序时（夏梦连等, 2018; Ma, 2021)。

虽然本研究的结果表明基于 NR-cRUM 的 CD-CAT 有很好的发展前景，但是这仍然没有充分发挥称名反应模型的优点。称名反应模型一个非常重要的优点是可以实现可修改答案的 CAT（Wang et al., 2017, 2019)。基于称名反应模型、可修改答案的 CD-CAT 值得进一步深入研究。

参考文献

陈平, 李珍, 辛涛. (2011). 认知诊断计算机化自适应测验的题库使用均匀性初探. *心理与行为研究, 9*(2), 125–132, 153.

高椿雷, 罗照盛, 郑蝉金, 喻晓锋, 彭亚风, 郭小军. (2017). CD-CAT 初始阶段项目选取方法. *心理科学, 40*(2), 485–491.

郭磊, 郑蝉金, 边玉芳, 宋乃庆, 夏凌翔. (2016). 认知诊断计算机化自适应测验中新的选题策略: 结合项目区分度指标. *心理学报, 48*(7), 903–914.

郭磊, 周文杰. (2021). 基于选项层面的认知诊断非参数方法. *心理学报, 53*(9), 1032–1043.

李佳, 丁树良, 况天昊. (2021). 区分度与测验进程相匹配的 CAT 选题策略. *江西师范大学学报(自然科学版), 45*(4), 384–389.

李瑜. (2014). *多选题认知诊断测验编制及多策略的多选题认知诊断模型的开发* (博士学位论文). 江西师范大学.

刘拓, 张佳慧, 辛涛. (2015). 多项选择题中干扰项信息的利用. *心理学探新, 35*(3), 261–265.

罗照盛, 杭丹丹, 秦春影, 喻晓锋. (2020). 可以处理补偿作用的认知诊断模型: CDINA 模型. *江西师范大学学报(自然科学版), 44*(5), 441–453.

罗照盛, 喻晓锋, 高椿雷, 李喻骏, 彭亚风, 王睿, 王钰彤. (2015). 基于属性掌握概率的认知诊断计算机化自适应测验选题策略. *心理学报, 47*(5), 679–688.

涂冬波, 蔡艳, 戴海琦. (2013). 认知诊断 CAT 选题策略及初始题选取方法. *心理科学, 36*(2), 469–474.

涂冬波, 郑蝉金, 戴步云, 汪文义. (2017). *计算机化自适应测验: 理论与方法*. 北京师范大学出版社.

王晓庆, 罗芬, 丁树良, 熊建华. (2016). 多级评分计算机化自适应测验动态调和平均选题策略. *心理学探新, 36*(3), 270–275.

夏梦连, 毛秀珍, 杨睿. (2018). 属性多级和项目多级评分的认知诊断模型. *江西师范大学学报(自然科学版), 42*(2), 134–138.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183.

Gao, X. L., Ma, W. C., Wang, D. X., Cai, Y., & Tu, D. B. (2021). A class of cognitive diagnosis models for polytomous data. *Journal of Educational and Behavioral Statistics, 46*(3), 297–322.

Gao, X. L., Wang, D. X., Cai, Y., & Tu, D. B. (2020). Cognitive diagnostic computerized adaptive testing for polytomously scored items. *Journal of Classification, 37*(3), 709–729.

Guo, L., & Zheng, C. J. (2019). Termination rules for variable-length CD-CAT from the information theory perspective. *Frontiers in Psychology, 10*, Article 1122. https://doi.org/10.3389/fpsyg.2019.01122

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262–277.

Henson, R., Roussos, L., Douglas, J., & He, X. M. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407–419.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Ma, W. C, & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253-275.

Ma, W. C. (2021). A higher-order cognitive diagnosis model with ordinal attributes for dichotomous response data. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2020.1860731

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19*(1), 91–100.

Templin, J., Henson, R., Rupp, A., Jang, E., & Ahmed, M.2008, March). *Cognitive diagnosis models for nominal response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models*. Springer.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035.

Wang, S. Y., Fellouris, G., & Chang, H.-H. (2017). Computerized adaptive testing that allows for response revision: Design and asymptotic theory. *Statistica Sinica, 27*(4), 1987–2010.

Wang, S. Y., Fellouris, G., & Chang, H.-H. (2019). Statistical foundations for computerized adaptive testing with response revision. *Psychometrika, 84*(2), 375–394.

Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively

based multiple-choice data. *Applied Psychological Measurement, 43*(5), 388–401.

Yu, X. F., Cheng, Y., & Chang, H.-H. (2019). Recent developments in cognitive diagnostic computerized adaptive testing (CD-CAT): A comprehensive review. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 307–331). Springer.

Zheng, C. J., & Chang, H.-H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*(8), 608–624.

Zheng, C. J., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 41*(7), 561–576.

A Comparative Study of Item Selection Methods in CD-CAT based on Nominal Response Model

Zhang Jie[1], *Luo Zhaosheng*[1], *Yu Xiaofeng*[1], *Qin Chunying*[2]
([1]School of Psychology, Jiangxi Normal University, Nanchang, 330022)
([2]School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032)

Abstract Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) combines the advantages of cognitive diagnosis and CAT, which could improve the efficiency and accuracy of CD-CAT. CD-CAT can be divided into two types: dichotomous and polytomous. Presently, the majority of researches on CD-CAT are based on dichotomous CD-CAT. However, among the practical tests in psychology and education, there are many polytomous items, which can be further divided into nominal polytomous and ordinal polytomous items according to whether there is an order or grade between every response category. Nominal polytomous items are items whose response categories are independent and without orders or grades between every response category. Although researchers have developed (ordinal) polytomous CDMs and corresponding CD-CAT, few nominal CDMs and CD-CAT are based on nominal responses.

This study introduces seven commonly used item selection methods in dichotomous CD-CAT into NCD-CAT (CD-CAT based on nominal response models). PMR (pattern match ratio) and test efficiency index are evaluated under different conditions between these item selection methods. Here are details of two simulation studies below. Study 1 compared the performance of NR_PWKL, NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, NR_MI, and NR_GDI methods under different test lengths (5, 10, 15, 20) and item pool qualities (high and low) in NCD-CAT. Results showed that: (1) the PMRs of NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, and NR_MI are higher than or equal to that of NR_PWKL, especially in short tests. (2) as test length gets longer, that PMR advantage is missing, which is the same as the results of Zheng and Chang (2016). (3) compared to test length, item quality has a greater impact on PMR. For instance, with item quality descending, the PMR declined about 30% among all conditions. Study 2 is an experiment on variable-length NCD-CAT that was conducted to compare the performance of each item selection method under the conditions of three maximum posterior probabilities (0.8, 0.85, 0.9) and two item qualities (high and low). The results showed that: (1)under all experimental conditions the average test lengths of NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, and NR_MI are shorter than that of NR_PWKL; the difference is more than 0.738. (2)affected by item quality, the average length of NR_GDI is smaller than that of NR_PWKL under high-quality conditions and larger than it under low-quality conditions.

To sum up, this study compared the performance of 7 commonly used item selection methods of dichotomous CD-CAT in NCD-CAT with different conditions (fixed and variable length). The simulation study showed that under most conditions, the NR_PWCDI, NR_PWACDI, NR_MPWKL, NR_SHE, and NR_MI methods performed well when compared to baseline algorithm NR_PWKL. This study has expanded the alternatives of item selection methods in NCD-CAT.

Keywords CD-CAT; nominal responses; NR-cRUM; item selection methods; multiple-choice items

附录 A

本研究中 NR-cRUM 模型是全模型 NR-DM（nominal response diagnostic model）的缩减形式。NR-DM 是 Templin 等（2008）基于 LCDM 和 LCA 模型，融入称名反应模型（NRM）思想而开发的称名反应认知诊断模型。其项目反应函数如下所示：

$$P(X=\boldsymbol{x})=\sum_{c=1}^{C}\eta_c\prod_{j=1}^{I}\left[\prod_{m_j\epsilon M_j}\pi_{j,c,m_j}^{I(x_j=m_j)}\right] \qquad (A-1)$$

其中，测验 Q 矩阵为 $Q=(q_{kj})_{K\times J}$，K 为属性个数，J 为项目个数，$C=2^K$ 表示 K 个属性把被试分成 C 个潜在类。π_{j,c,m_j} 表示对于项目 j，第 c 类被试选择选项 m_j 的概率，且有 $\sum_{m_j\epsilon M_j}\pi_{j,c,m_j}=1,\ \forall j,c.$ η_c 是第 c 类被试在被试群体所占的比例，且 $\sum_{c=1}^{2^K}\eta_c=1.$ M_j 是项目 j 所有可能的选项，这里的 $\boldsymbol{x}$ 是反应向量，$\boldsymbol{x}=(x_1,\dots,x_J)$。$I(\cdot)$ 是一个指示函数，当 $x_j=m_j$ 时，它的值为 1，当 $x_j\neq m_j$ 时，它的值为 0. 下面是 π_{j,c,m_j} 的参数化形式：

$$\pi_{j,c,m_j}=P\left(X_j=m_j\middle|\boldsymbol{\alpha}_c\right)=\frac{exp\left(\lambda_{0,j,m_j}+\boldsymbol{\lambda}_{j,m_j}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_j)\right)}{\sum_{m_j\epsilon M_j}exp\left(\lambda_{0,j,m_j}+\boldsymbol{\lambda}_{j,m_j}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_j)\right)} \qquad (A-2)$$

$$\boldsymbol{\lambda}_{j,m_j}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_c,\boldsymbol{q}_j)=\sum_{k=1}^{K}\lambda_{1,j,k,m_j}(\alpha_{c,k}q_{j,k})+\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\lambda_{2,j,k,l,m_j}(\alpha_{c,k}\alpha_{c,l}q_{j,k}q_{j,l})+\cdots \qquad (A-3)$$

$\alpha_c=(\alpha_{c1},\dots,\alpha_{cK})'$ 是第 c 类被试的属性掌握模式。对于项目 j 的每个选项都有三类参数，分别是：（1）截距参数 λ_{0,j,m_j}；（2）属性 k 的主效应参数 λ_{1,j,k,m_j}；（3）属性 k 与属性 l 的交互效应参数 λ_{2,j,k,l,m_j}，以及两个以上属性的交互效应参数。为了模型的可识别性，需要约束各类参数之和为 0，如：$\sum_{m_j\epsilon M_j}\lambda_{0,j,m_j}=0,\ \forall j;\ \sum_{m_j\epsilon M_j}\lambda_{1,j,k,m_j}=0,\ \forall j,k$ 等。

由于 NR-DM 是基于 LCDM 模型的拓展，因此它是一个更一般、更广义的称名反应认知诊断模型。在 NR-DM 中，若仅考虑题目考察的所有属性的交互效应，则该模型等价于 MC-DINA 模型，是一个非补偿模型；若不考虑题目考察的所有属性的交互效应，则该模型属于补偿模型。在实际应用中，考虑了交互效应的 NR-DM 需要估计的参数非常多，需要非常大的样本量才能够进行准确的参数估计，所以 NR-cRUM 模型是一个更实际的选择，具有较大的实用性（Templin et al., 2008）。NR-cRUM 模型的项目反应函数如下所示：

$$\pi_{j,c,m_j}=P\left(X_j=m_j\middle|\boldsymbol{\alpha}_c\right)=\frac{exp\left(\lambda_{0,j,m_j}+\sum_{k=1}^{K}\lambda_{1,j,k,m_j}(\alpha_{c,k}q_{j,k})\right)}{\sum_{m_j\epsilon M_j}exp\left(\lambda_{0,j,m_j}+\sum_{k=1}^{K}\lambda_{1,j,k,m_j}(\alpha_{c,k}q_{j,k})\right)} \qquad (A-4)$$

其中，$\sum_{m_j\epsilon M_j}\lambda_{0,j,m_j}=0,\ \forall j;\ \sum_{m_j\epsilon M_j}\lambda_{1,j,k,m_j}=0,\ \forall j,k.$ 且如果 m_j 是项目 j 的正确答案选项，则 $\lambda_{1,j,k,m_j}>0,\ \forall k.$

根据上面的项目反应函数可以看出，在全模型和缩减模型中，属性定义在选项水平上，属性的定义更精细。且项目参数定义在题目和选项的交互水平上，模型参数的可解释性更强。

为了更好地理解 NR-cRUM 模型，不妨借助下面这个数学推理题来理解（Templin et al., 2008）。

甲住在距离乙家 $\frac{2}{3}$ 米的地方，甲从家出发朝着乙家走去，某时刻甲距离乙家还有 $\frac{1}{4}$ 米，问：此时甲已经走了多少路程？下面哪个选项表示甲走过的路程。（注：各选项的线段长度

为 1 米）

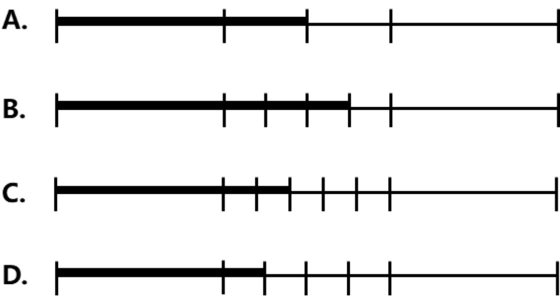


图 A-1

这道题考察了两个属性：1. 识别一个指定量的参照单位；2. 分数减法。

假定该题目各选项的参数如表 A-1 所示，根据公式（A-4）可算出不同属性掌握模式的被试选择不同选项的概率，如图 A-2 所示（该题正确选项为 D）。

表 A-1 NR-cRUM 模型题目各选项的参数

	A	B	C	D
λ_{0,j,m_j}	1	0.5	0.5	-2
$\lambda_{1,j,1,m_j}$	-1	-1	0	2
$\lambda_{1,j,2,m_j}$	-1	0	-1	2

表 A-2 NR-cRUM 模型不同 KS 的被试选择各选项的概率

	A	B	C	D
(0, 0)	0.442	0.268	0.268	0.022
(0, 1)	0.235	0.387	0.143	0.235
(1, 0)	0.235	0.143	0.387	0.235
(1, 1)	0.041	0.068	0.068	0.824

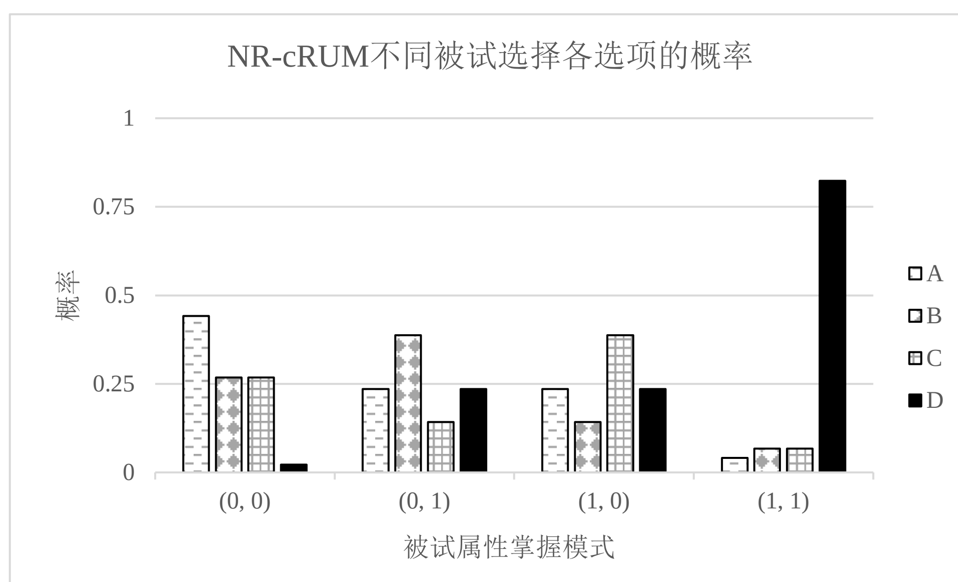


图 A-2 NR-cRUM 模型下不同 KS 的被试选择各选项的概率

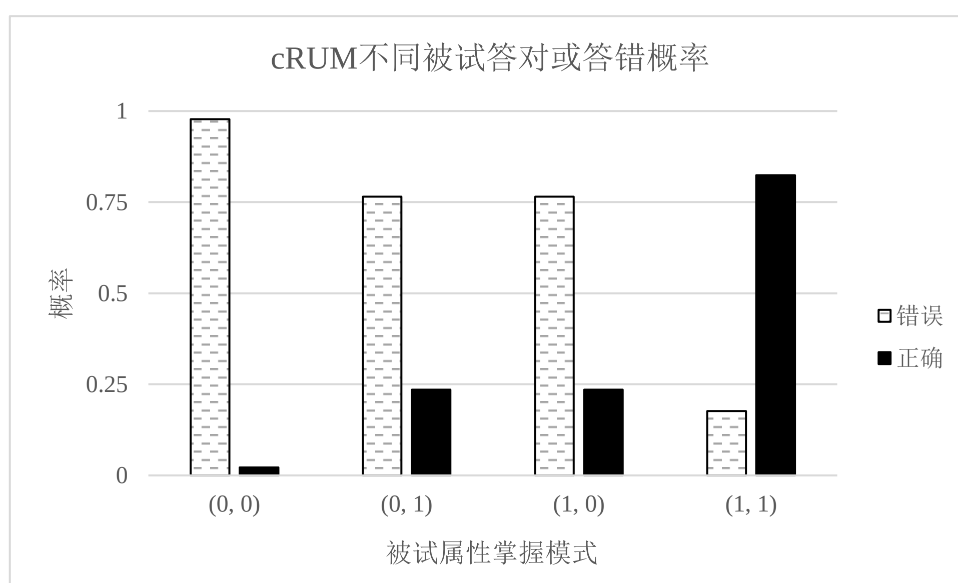


图 A-3 cRUM 模型下不同 KS 的被试答对或答错的概率

经过上面这个例子，可以看出 NR-cRUM 模型对被试选择每个选项的概率都进行了参数化，每个选项都有一套参数，根据公式（A-4）可以计算出被试选择每个选项的概率。结合表 A-2 和图 A-2 可知，两个属性都掌握了的被试（1, 1）有 0.824 的概率会选择正确选项 D，而被试（0, 0），(0, 1），(1, 0）分别有较大概率选择 A, B, C 选项，即被试会选择与其能力相“匹配”的选项，所以说该模型能够利用每个选项的信息，从而能用更少的题目达到对被试能力更准确的测量。相较于传统 0-1 计分的 CDM（cRUM），对比图 A-2 和图 A-3 可知，NR-cRUM 模型可以根据被试选择不同选项，区分出被试属于四种 KS 中的哪一种，而在 cRUM 模型中，只能将（1, 1）与其它 KS 区分出来。很显然，通过对不正确选项的建模，NR-cRUM 模型可以最大限度地获取选择题的诊断信息，并以此提高属性掌握模式的估计准确性和效率。

附录 B

Q 矩阵

$$
\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1
\end{pmatrix}
$$

附录 C

评价指标

判准率指标有属性判准率（attribute match ratio, AMR）和模式判准率（pattern match ratio, PMR），计算公式如（C-1），（C-2）。

$$AMR_k = \frac{\sum_{i=1}^{N} I(\hat{\alpha}_{ik}, \alpha_{ik})}{N}, (k = 1,2, \dots, K) \quad (C-1)$$

$$PMR = \frac{\sum_{i=1}^{N} I(\hat{\boldsymbol{\alpha}}_i, \boldsymbol{\alpha}_i)}{N} \quad (C-2)$$

其中 K 表示属性总个数，N 表示被试人数，$\hat{\alpha}_{ik}$ 和 α_{ik} 分别是 $\hat{\boldsymbol{\alpha}}_i$ 和 $\boldsymbol{\alpha}_i$ 的第 K 个元素。若 $\hat{\alpha}_{ik} = \alpha_{ik}$，则称判准属性 k 一次，AMR 的值越大，说明对单个属性的判准率越高。$\hat{\boldsymbol{\alpha}}_i$ 和 $\boldsymbol{\alpha}_i$分别是第 i 个被试的 KS 估计值和 KS 真值，若 $\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$ 时，则称判准被试的 KS 一次，PMR 的值越大，说明对被试整个属性掌握模式的判准率越高。

题库安全性指标主要有 χ^2 和测验重叠率（test overlap rate, TOR）。χ^2 指标描述的是理想项目曝光率分布与真实项目曝光率分布之间的差异，其计算公式如下：

$$\chi^2 = \frac{\sum_{j=1}^{J} \left(er_j - {}^{L}/_{J}\right)^2}{\left({}^{L}/_{J}\right)} \quad (C-3)$$

其中 J 是题库大小，L 是测验的长度，L/J 是均匀分布的项目曝光率。er_j 是第 j 个题目的曝光率。χ^2 指标越小越好，说明题库使用越均匀。测验重叠率反映的是不同被试调用相同题目的比例，其定义为两个随机抽取的被试作答相同题目的期望数除以题长，计算公式如下：

$$TOR = \frac{\sum_{j=1}^{J} T_j \times (T_j - 1)}{L \times N \times (N-1)} \quad (C-4)$$

其中，T_j 是第 j 题被调用的次数。测验重叠率也是越小越好（陈平 等, 2011）。

附录 D

表 D-1 七种选题方法的模式判准率、测验重叠率和卡方值

题长	方法 (NR_)	高质量			低质量		
		PMR	TOR	χ^2	PMR	TOR	χ^2
5	PWKL	0.624	0.121	67.483	0.335	0.174	99.682
	PWCDI	0.711	0.228	132.12	0.367	0.339	198.729
	PWACDI	0.714	0.207	119.071	**0.373**	0.305	178.192
	MPWKL	0.704	0.24	139.379	0.357	0.36	211.268
	SHE	**0.736**	0.206	119.004	0.352	0.317	185.327
	MI	**0.736**	0.206	119.004	0.352	0.317	185.327
	GDI	0.672	0.434	255.219	0.322	0.307	179.57
10	PWKL	0.917	0.229	127.318	0.621	0.274	154.604
	PWCDI	0.948	0.291	164.595	0.643	0.368	211.092
	PWACDI	0.948	0.282	159.451	0.633	0.341	194.806
	MPWKL	0.948	0.296	167.637	0.641	0.386	221.625
	SHE	**0.951**	0.23	128.057	**0.646**	0.335	190.999
	MI	**0.951**	0.23	128.057	**0.646**	0.335	190.999
	GDI	0.948	0.426	245.786	0.607	0.401	230.914
15	PWKL	0.988	0.269	146.449	0.789	0.324	179.775
	PWCDI	0.993	0.345	191.841	0.798	0.381	213.913
	PWACDI	0.993	0.34	189.126	**0.808**	0.365	204.257
	MPWKL	**0.994**	0.346	192.899	0.803	0.393	221.121
	SHE	0.993	0.256	138.986	**0.808**	0.342	190.392
	MI	0.993	0.256	138.986	**0.808**	0.342	190.392
	GDI	0.993	0.417	235.127	0.772	0.435	245.966
20	PWKL	0.999	0.291	154.467	0.888	0.359	195.715
	PWCDI	0.999	0.379	207.723	0.892	0.398	219.117
	PWACDI	0.999	0.378	206.622	0.897	0.389	213.382
	MPWKL	**1**	0.381	208.528	0.895	0.407	224.16
	SHE	0.999	0.273	143.646	**0.902**	0.355	192.85
	MI	0.999	0.273	143.646	**0.902**	0.355	192.85
	GDI	0.998	0.425	235.231	0.874	0.451	250.801

表 D-2 各选题方法的判准率和题长的描述统计量

终止规则	方法 (NR_)	高质量					低质量				
		PMR	Min	Max	Mean	SD	PMR	Min	Max	Mean	SD
0.8	PWKL	0.851	4	20	7.108	1.826	0.811	6	20	13.304	3.741
	PWCDI	0.853	4	18	6.249	1.515	0.813	6	20	**12.99**	3.706
	PWACDI	0.857	4	14	6.251	1.509	0.828	6	20	13.033	3.756
	MPWKL	0.856	4	16	6.262	1.516	0.824	6	20	13.005	3.721
	SHE	0.852	4	15	**6.119**	1.547	0.823	6	20	13.052	3.709
	MI	0.852	4	15	**6.119**	1.547	0.823	6	20	13.052	3.709
	GDI	0.866	4	14	6.492	1.32	0.803	6	20	13.833	3.77
0.85	PWKL	0.894	4	20	7.719	1.991	0.841	6	20	14.473	3.75
	PWCDI	0.891	4	19	6.822	1.664	0.847	7	20	14.189	3.711
	PWACDI	0.889	4	15	6.823	1.678	0.857	7	20	14.188	3.711
	MPWKL	0.892	4	18	6.83	1.673	0.856	6	20	**14.169**	3.716
	SHE	0.883	4	17	**6.766**	1.696	0.857	7	20	14.279	3.734
	MI	0.883	4	17	**6.766**	1.696	0.857	7	20	14.279	3.734
	GDI	0.896	4	15	7.056	1.433	0.833	7	20	15.031	3.687
0.9	PWKL	0.932	5	20	8.393	2.101	0.866	7	20	15.872	3.536
	PWCDI	0.926	5	19	**7.549**	1.831	0.869	7	20	15.632	3.564
	PWACDI	0.927	5	19	7.585	1.862	0.874	7	20	15.62	3.56
	MPWKL	0.928	5	19	7.563	1.833	0.874	8	20	**15.598**	3.569
	SHE	0.922	5	18	7.655	1.82	0.881	7	20	15.724	3.56
	MI	0.922	5	18	7.655	1.82	0.881	7	20	15.724	3.56
	GDI	0.927	5	17	7.884	1.594	0.856	7	20	16.377	3.424

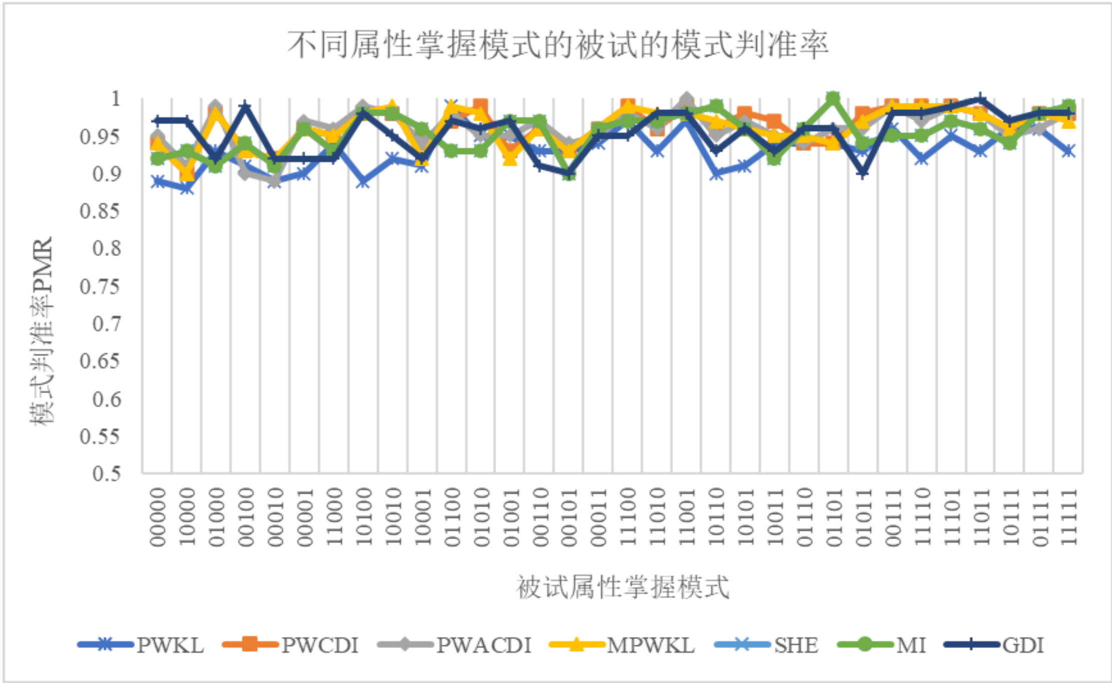


图 D-1 高质量题目（10 题）条件下不同 KS 被试的 PMR

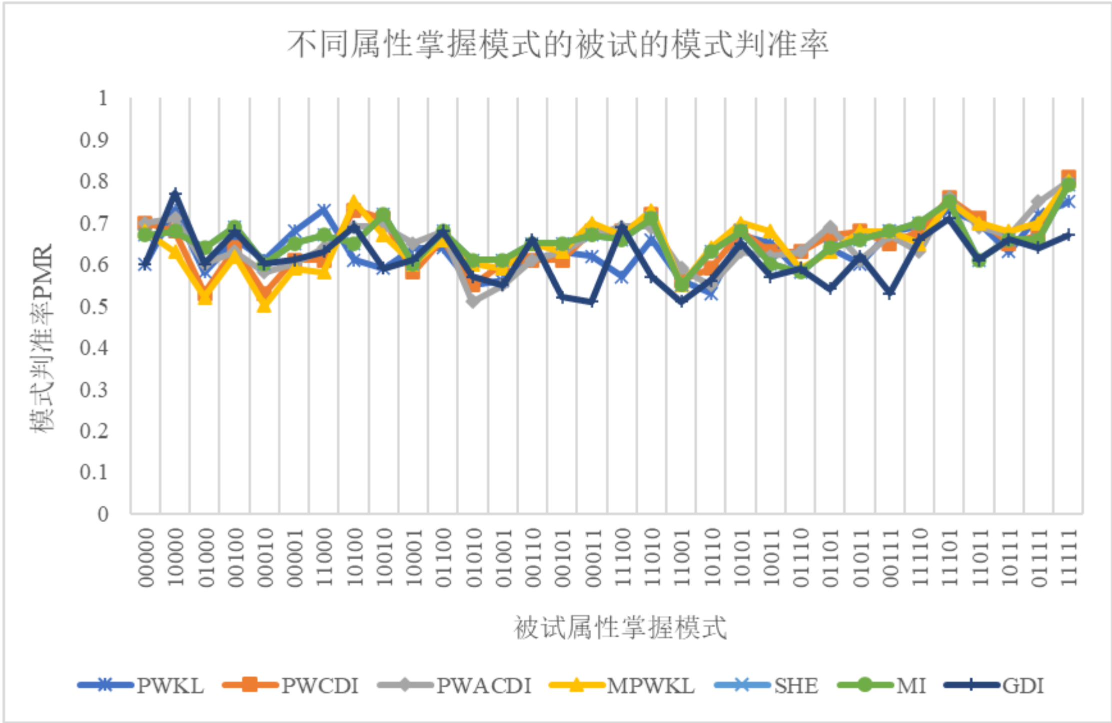


图 D-2 低质量题目（10 题）条件下不同 KS 被试的 PMR

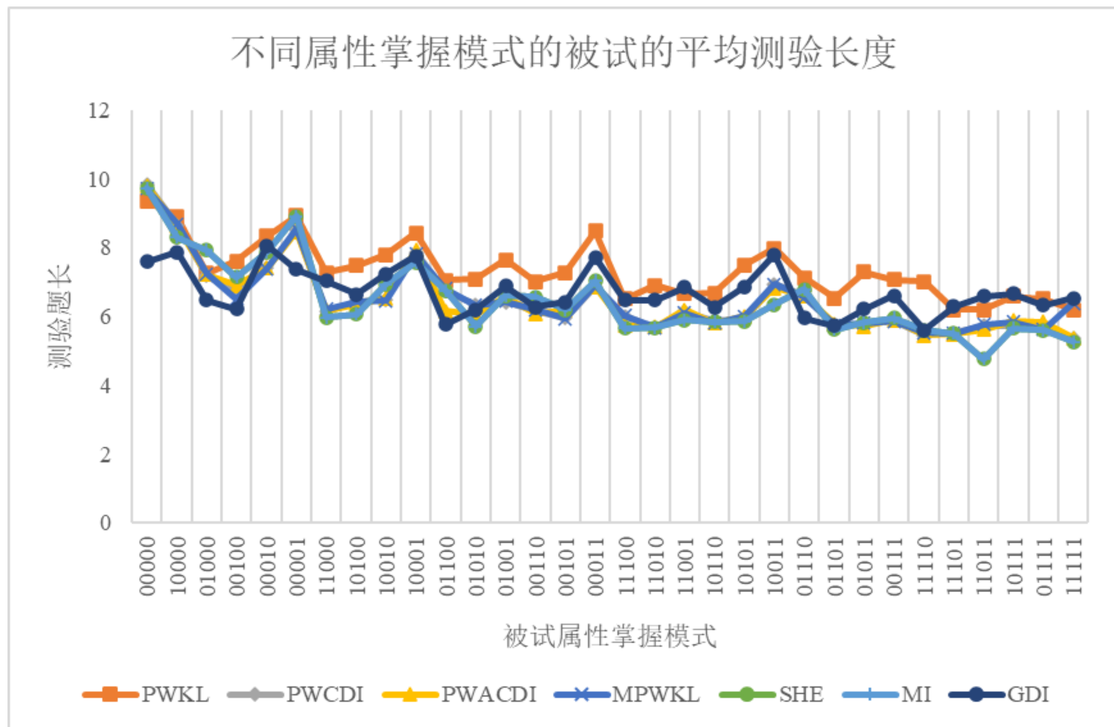


图 D-3 高题目质量和最大后验概率 0.85 条件下的平均题长

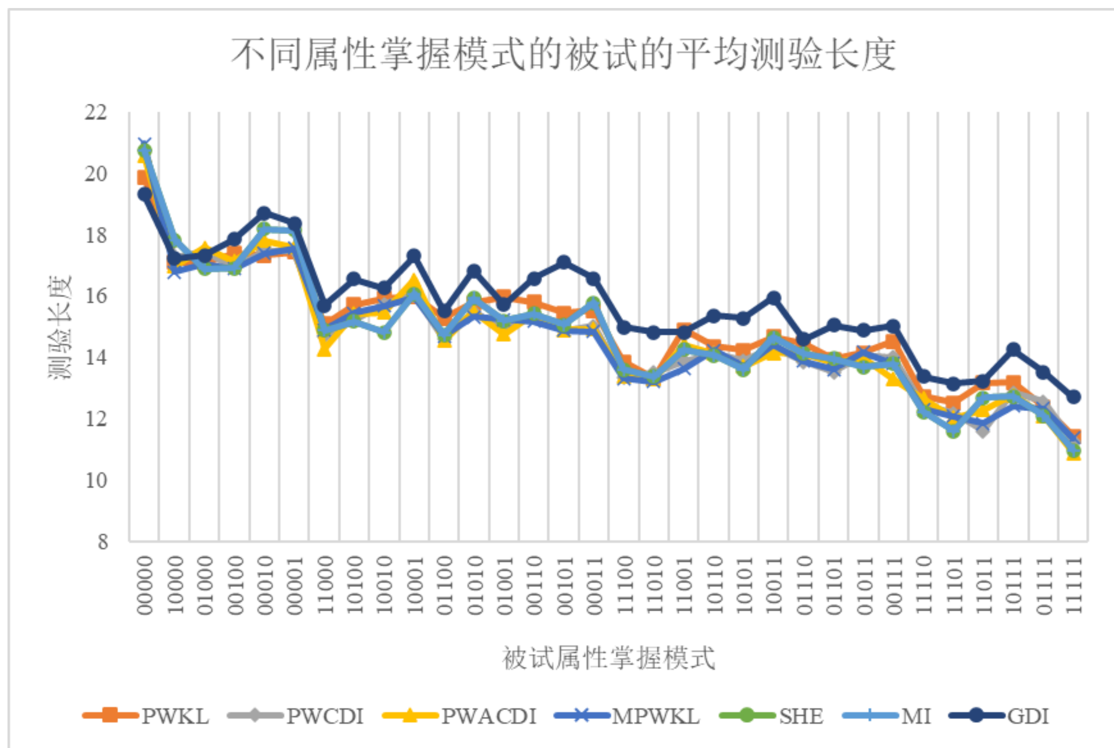


图 D-4 低题目质量和最大后验概率 0.85 条件下的平均题长